# 基于LDA模型的国内档案学热点主题及演化研究*

周洁[1]　　盛梅[2]

[1] 宁波大学档案馆　　宁波　315211 [2] 浙江省立同德医院　杭州　310000

**摘要：［目的/意义］**采用LDA模型发现近年来我国档案学的研究热点和发展趋势，为我国档案学主题研究提供数据支撑和参考价值。**［方法/过程］**选取2012年-2022年间9本档案学核心期刊的中文摘要为分析样本，以中国知网数据库（CNKI）为来源数据库，通过Python的开源工具包pkuseg进行中文分词，gensim搭建LDA模型，pyLDAvis将各主题基于web的交互式可视化。根据pyLDAvis可视化结果为主题命名，根据文档－主题概率分布情况并结合时间项分析热点主题和主题演化过程。**［结果/结论］**根据LDA模型能够有效地区分国内档案学领域研究的主题。2012-2022年国内档案学领域有14个主题，其中热点主题有5个；3个主题呈上升趋势，1个主题呈下降趋势，10个主题呈不同程度的波段趋势。

**关键词：**档案学　LDA模型　热点主题　主题演化

**分类号：**G270

**作者简介：**周洁，馆员，硕士，E-mail：zhoujie1@nbu.edu.cn；盛梅，馆员，硕士。

## 1 引言

档案学领域中的文献研究法通过各种横向和纵向的比较为档案学者梳理出清晰的发展脉络，为其快速了解档案学领域提供有效便捷的途径，为更深入的后续研究提供扎实的数据支撑。文献研究法主要分为两种方式：文献计量法和内容分析法。

文献计量法主要是对文献的各类外部特征进行研究分析。如；宋进[1]等根据地域特征，对河南省十八个地区的公共档案馆发表的学术论文分布情况进行分析，多角度、多层次、全面、客观地反映河南省公共档案馆科学研究的态势,进一步揭示河南省公共档案馆研究人员的科研创新能力；马双双[2]等根据研究机构特征，从整体合作网络与核心合作网络两个层次对我国档案学领域研究机构合作网络进行了分析，总结出核心机构网络连通性较好,但核心机构合作的多样性有待加强的结论；李英[3]根据基金项目特征，从立项数量、立项类型、立项单位及其所属行业、项目负责人以及项目的主题内容特征等方面进行系统分析比较，预测未来的发展趋势；陶俊[4]等根据高被引文特征，从主题、演化和引用结构方面对档案学CSSCI来源期刊近年的高被引论文统计分析,刻画档案学科结构进而评价总体竞争力；杨万欢[5]等根据期刊分布、核心作者、项目基金、研究方法、研究内容论文类型等特征，对国内智慧档案馆研究的学术论文进行统计分析，概括当前国内智慧档案馆态势,并对未来发展和研究导向提出展望。

内容分析法，顾名思义，是指直接对文献的文字内容进行分析。目前，内容分析法主要分为三种方式：人工分析，档案学者通过对文献的大量阅读与理解，进行综合分析与总结；词频分析，通过统计各词语出现的频率，对高频词、共词等进行研究；文本分析，通过各类先进的文本模型算法挖掘文本的主题，语义等。人工分析方式如于欢欢6等梳理档案领域中关于区域块链技术的文献，将其分为:区块链技术应用的可行性，面向专门档案领域的应用，基于区块链技术的档案管理平台，区块链技术应用于档案领域的动力因素，阻力因素和推广策略

chinaXiv:202302.00055v1

六大主题。词频分析方式如张晓培[7]通过词频分析和建立的高频词可视化共词网络图，得出档案信息、档案开放、现行文件、国家档案馆、信息查询、档案工作、档案利用是当前政府信息公开与档案相关研究领域研究的重点与热点。文本分析方式如逯万辉[8]等利用 ATM 主题模型计算作者所属主题分布情况，构建作者主题内合作网络及跨主题合作网络并测度不同网络内学者的中心度，以此来反映作者在研究领域的研究专业性和知识创新性；宋雪雁[9]等利用 RDF 数据模型存储清代祭祀礼器知识，构建清代祭祀礼器知识图谱，通过检索知识图谱进行知识发现；马海群[10]等利用 LDA 主题聚类及相似度计算方法对《中华人民共和国档案法》（2020 修订版，以下简称档案法）和《“十四五”全国档案事业发展规划》进行主题和内容协同关系研究，发现其在档案信息化建设、档案人才培养方面具有较强的协同性。

文献计量法最为简单和便捷，能够对大量期刊论文进行快速分析，但其缺少对文献主要内容的理解。人工分析方法最为准确和精准，但需要大量阅读相关文献并做好总结归纳，耗时久的同时还需要作者有很强的逻辑分析能力。词频分析方法没有考虑词频与文档，文档与时间序列之间的关联关系，很难全面呈现文献所表达的内涵。文本分析方法通过搭建各类文本模型，使用计算机辅助运算能够较快发掘文献的深层含义。通过知网检索发现，基于文本分析方法研究档案学的文献较少，值得进一步探索与深究。

Griffiths[11]等于 2004 年首先运用 Gibbs 抽样算法来推断 LDA 模型，并用于提取文献主题。经过学者门后续不断地改进与扩展，是目前文本模型中最常用的模型之一。基于关键字的 LDA 模型会出现主题提取不全情况，基于全文的 LDA 模型因数据量过大存在“噪音”干扰。而摘要用简短的语句总结概括了整个文献的主要内容，是作者思想的精华提炼。故本文拟采用 LDA 模型对我国档案学 2012-2022 年这 11 年的 9 本档案学核心期刊的中文摘要进行文本分析，通过 Python 语言的的开源工具 pkuseg 包对文献摘要中文分词，gensim 包在分词基础上搭建 LDA 模型，运用困惑度来确定档案学主题模型数量，pyLDAvis 包将各主题基于 web 的交互式可视化。最后，根据可视化网页命名各主题，根据文档一主题概率分布情况并结合时间项分析热点主题和主题演化过程。

## 2 数据来源及预处理

### 2.1 数据来源

本文将中国知网数据库作为数据源，从北大核心期刊目录中选定与档案学紧密相关的 9 本核心期刊作为期刊来源。通过中国知网的高级检索功能，采用文献来源=“档案学通讯”或“档案学研究”或“中国档案”或“档案管理”或“档案与建设”或“历史档案”或“民国档案”或“北京档案”或“浙江档案”，发表时间范围为 2012-2022 年进行组合查询。除去新闻动态、会议培训通知、上级部门传真等与研究不相关的文献，以及无摘要内容的文献，共计 11561 篇。通过查新（引文格式）导出 excel 表作为实验数据表。实验数据表中期刊来源和时间分布情况如表 1 所示。

表 1　期刊时间分布表

| | 北京档案 | 档案管理 | 档案学通讯 | 档案学研究 | 档案与建设 | 历史档案 | 民国档案 | 浙江档案 | 中国档案 |
|---|---|---|---|---|---|---|---|---|---|
| 2012 | 138 | 134 | 144 | 111 | 128 | 44 | 54 | 70 | 39 |
| 2013 | 137 | 128 | 139 | 117 | 125 | 53 | 60 | 100 | 58 |

chinaXiv:202302.00055v1

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 2014 | 113 | 128 | 131 | 108 | 130 | 41 | 58 | 128 | 69 |
| 2015 | 126 | 117 | 124 | 135 | 137 | 40 | 56 | 124 | 78 |
| 2016 | 135 | 173 | 120 | 134 | 144 | 35 | 49 | 133 | 76 |
| 2017 | 120 | 162 | 123 | 167 | 150 | 35 | 54 | 183 | 77 |
| 2018 | 119 | 161 | 127 | 131 | 143 | 43 | 42 | 182 | 52 |
| 2019 | 128 | 163 | 92 | 119 | 190 | 48 | 40 | 191 | 61 |
| 2020 | 157 | 275 | 84 | 121 | 221 | 54 | 46 | 203 | 118 |
| 2021 | 171 | 296 | 81 | 124 | 233 | 51 | 45 | 211 | 159 |
| 2022 | 158 | 222 | 80 | 117 | 245 | 44 | 29 | 151 | 141 |
| 合计 | 1502 | 1959 | 1245 | 1384 | 1846 | 488 | 533 | 1676 | 928 |

### 2.2 数据预处理

本文选择 pkuseg 包进行中文分词，pkuseg 包是北京大学语言计算与机器学习研究组罗睿轩[12]等开发的一款多领域中文分词工具包，其简单易用，支持细分领域分词，有效提升了分词准确度。pkuseg 包可自定义词典，如将档案专有名词“智慧档案馆”作为一个词处理，而不是拆分成“智慧”、“档案馆”两个词，本文提取实验数据表中的关键词作为自定义词典。在分词的同时，pkuseg 包支持停用词过滤。中文停用词表（cn_stopwords）包含通用的无意义词表，但未包含“本文”、“阐述”等论文结构的无意义词，本文选择中文停用词表+“人工补著”的方式设置停用词表。

### 2.3 最优主题数选择

经过数据预处理后得到文档-词项文件，将其作为 LDA 模型的输入源。使用 gensim 包构建 LDA 模型。目前，LDA 模型最优主题数[13]的确定主要有困惑度、一致性、主体间相似度、经验法四种方式。为了运算方便，本文采用困惑度来确定最优主题数，实验结果如图 1 所示。当主题数目为 14 时，困惑度较小，且随着后续主题数量的增加，困惑度值基本维持不变。综上所述，设置主题个数 K=14，文档-主题分布的先验信息 alpha=50/k，主题-词项分布的先验信息 eta=0.01，各主题下最相关的词项数量 num_words=30。

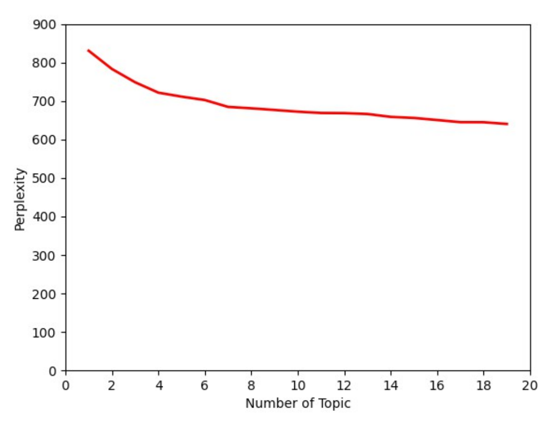



图 1 主题数与困惑度变化分布图

## 3. 实验结果与分析

### 3.1 实验结果

chinaXiv:202302.00055v1

chinaXiv:202302.00055v1

经过 LDA 模型测试后，得到词项-主题概率分布文件、文档-主题概率分布文件。使用 pyLDAvis 包将 LDA 模型运算结果保存为交互式 HTML 文件，如图 2 所示。pyLDAvis 包主要由 Carson Sievert[14]等提供，通过交互式网页的演示，帮助用户快速直接的观察各主题情况。

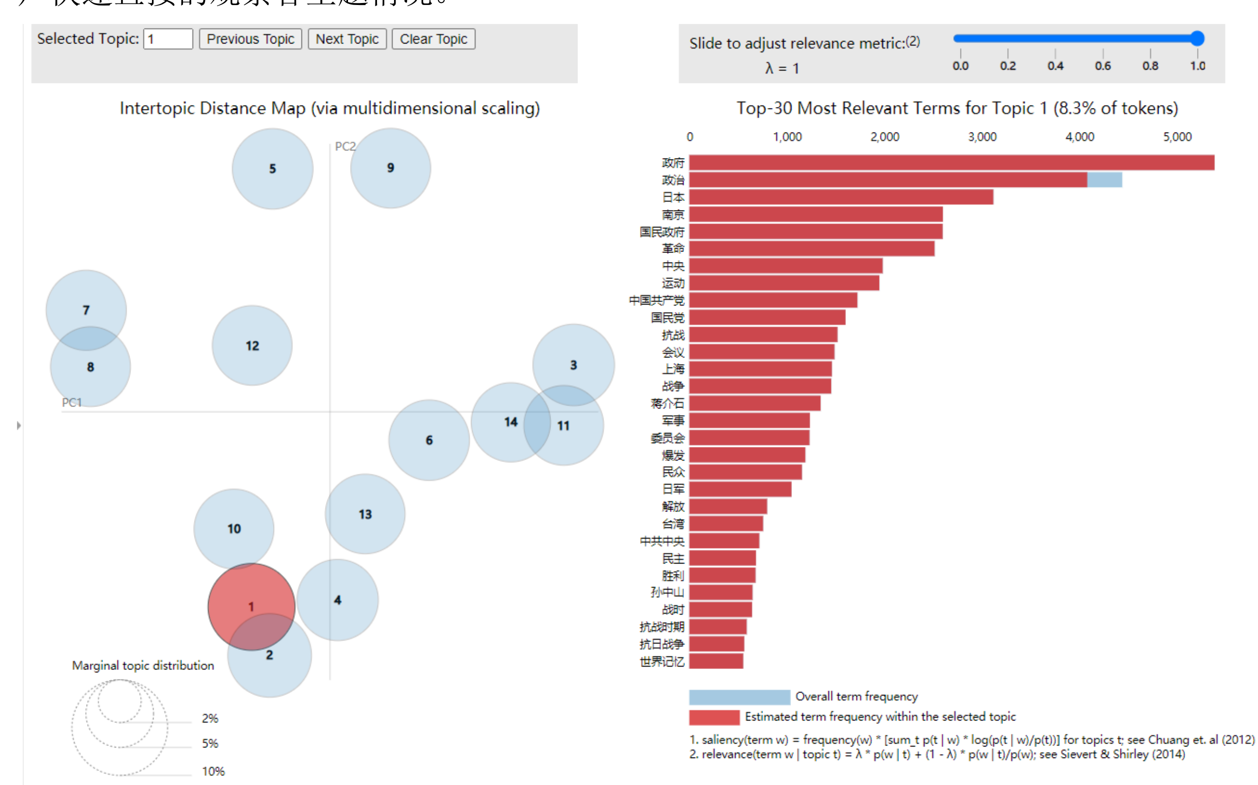



图 2　LDA 模型可视化结果图

图 2 左侧的蓝色圆圈表示 LDA 模型的 14 个主题，蓝色圆圈之间的距离表示主题间的相似性，圆圈间的交叉表示两个主题的特征词有交叉。右侧目前显示左侧编号为 1 的主题的前 30 个特征词项，每个词项的蓝色部分代表其在整个文档中所占权重，红色部分代表其在该主题中所占权重。右上角 λ 值可从 0~1 之间调节，λ 值越接近 1，表示在该主题下出现越频繁的词项与该主题更相关；λ 值越接近 0，表示在该主题下越专有的词项与主题更相关。图 2 表明 14 个主题分布较为均匀合理，大部分特征词仅归属于一个主题。根据图 2 可视化结果对各主题进行命名，主题与排名靠前的 10 相关词项如图表 2 所示。

表 2　主题-词项分布表

| 主题编号 | 主题命名 | 词项 |
|---|---|---|
| Topic1 | 民国档案 | 政府、政治、日本、南京、国民政府、革命、中央、运动、中国共产党、国民党 |
| Topic2 | 历史档案 | 历史、清代、史料、明代、历史档案馆、珍贵、清朝、档案史料、中国古代、盛京 |
| Topic3 | 档案服务 | 服务、利用、共享、档案利用、用户、利用服务、知识服务、资源整合、用户需求、资源共享 |
| Topic4 | 档案价值 | 档案、价值、文化、记录、文化建设、档案文化、档案价值、文字、符号、承载 |
| Topic5 | 档案法律 | 档案法、原则、档案行政、规定、法律、责任、立法、监督、程序、修订 |
| Topic6 | 档案信息化 | 技术、系统、档案数据、信息化、信息技术、档案信息化、信息化建设、智能化、智慧档案馆、顶层设计 |

| | | |
|---|---|---|
| Topic7 | 文献研究 | 文献、主题、来源、分布、数据库、检索、统计、期刊、档案领域、热点 |
| Topic8 | 档案规范 | 规范、标准、监管、行业、档案行业、公文、规范化、格式、解读、编制 |
| Topic9 | 企业档案 | 项目、企业档案、工程、公司、生产、材料、建设项目、项目档案、集团、工程档案 |
| Topic10 | 人才培养 | 能力、专业、培养、档案工作者、人才、档案专业、素质、职业、培育、培训 |
| Topic11 | 电子档案 | 电子文件、电子档案、流程、真实性、元数据、完整性、试点、安全性、一体化、管理系统 |
| Topic12 | 档案学理论 | 理论、档案学、学术、学科、档案学研究、理论研究、我国档案学、中国档案学、基础理论、档案学理论 |
| Topic13 | 档案开发 | 开发、档案文化、产品、记忆、宣传、编纂、挖掘、档案编研、口述历史、媒体 |
| Topic14 | 档案管理 | 管理、档案管理、档案整理、收集、档案收集、业务、档案库房、分类、接收、移交 |

3.2 主题分析

下面对 14 个主题进行简要分析，并根据文档-主题概率分布文件，选取各主题下概率较高的文档进行阐述。

（1）民国档案。民国档案是指从 1912 年辛亥革命爆发到 1949 年中华人民共和国成立期间的史料档案，是对民国期间的革命战争、国民政府建设等剖析。如张展[15]考察华北、华中日军围绕“新中央政府”之争所产生的龃龉，探讨日军对伪政权政策的出台背景与模式。姚江鸿[16]探讨 1944 年国民党实施改革的原因、动力以及内部决策经过，以此管窥国民党内部复杂的权力结构、人事关系，以及蒋介石在用人和处理战略危机时的一些思维特征。

（2）历史档案。历史档案以明清时期的史料档案为主，主要是对名人名事、官僚制度、图书著作等进行研究。如李兵[17]利用《清实录》、《钦定科场条例》等文献，详细描述了湖南士子呼吁分闱、湖南巡抚奏请分闱，实现分闱：新建与重修贡院的整个过程；陈晨[18]梳理盛京巡察官的发展脉络,检视其职掌与权力的演进过程及实际运作；赵彦昌[19]概括了清代盛京总管内务府衙门处理皇室事务往来公文的副本档《黑图档》中有关凤凰楼的信息，用以探究凤凰楼职能的变化、日常维护与修缮事务。

（3）档案服务。档案服务是指档案部门根据用户实际需求，探索新平台、新方法、新模式等，以期通过资源整合和共享，为用户提供更加个性化、智能化服务。如王成兴[20]等分别对档案信息服务平台的概念和建设必要性进行了界定与分析，并对档案信息服务交互平台模型进行了构建；傅永珍[21]通过分析档案信息用户的多样性、个性化以及档案信息用户需求的个性化，探讨了面向用户需求的档案信息个性化服务；连志英[22]基于国内外档案机构在社会化媒体平台环境下档案信息服务的研究现状，构建了参与式档案信息服务模式。

（4）档案价值。档案价值主要以理论分析为主，档案作为文字的原始记录，承载了丰富的历史数据，为人类提供了文化价值，情感价值、凭证价值、经验价值、经济价值等多种价值。如赵爱学[23]从档案载体“物化”价值、承载记忆符号的“文化价值”两个方面进行考察，探讨档案的文化起源。通过对档案包含着人与自身自我意识的关系、人与自然的物质变量的关系、人与社会行为的关系的历史考察，揭示档案的文化属性；杨光[24]等从话语形式、主题认知和文化面貌三个方面分析了档案从文字（档案文本）、图像（档案影像）到二进制代码（数字档案）的变迁；管先海[25]深入探讨了档案价值、档案主体价值以及档案客体价值。

chinaXiv:202302.00055v1

（5）档案法律。档案法律的研究主要围绕档案法展开，包括对法律的解读、建议、实施措施等。如梅帅[26]在新《中华人民共和国行政处罚法》视域下，发现我国新修订的档案法在立法规定、立法技术、制度建设、监督执行等方面存在一定的滞后与缺陷，建议更新完善《档案行政处罚程序暂行规定》、改进档案行政处罚程序的制定技术、完善档案行政处罚程序相关制度、健全档案行政处罚监督执行规定。

（6）档案信息化。档案信息化是指通过体系建设、技术引入等实现对档案数字资源的安全、长久管理。研究以数字档案馆建设为基础，期望经过不断打磨，最终发展为智慧档案馆。如金波[27]等总结了档案数据安全风险种类、成因及特点，从加强档案数据安全法治建设、推进档案数据安全协同共治、打造档案数据安全技术高地、培育档案数据安全专业人才四个层面分析档案数据安全保障路径；赵湘渝[28]描述智能化技术在档案馆建筑、馆库、安防、测控、管理、建设、生态智能化方面的应用现状，依据现有应用中“五定五不定”的特点，从定位、认识、研究、行动四个方面提出建议。

（7）文献研究。文献研究是指学者们对档案学文献期刊等进行研究分析，从宏观和微观各个角度总结和概括档案学领域的研究结果。如王国强[29]从年度、期刊、作者、机构分布对 1985-2018 年在档案学 10 本期刊上的有关家庭建档的 79 篇研究成果进行分析，对档案行政管理部门如何施政研究不够、家庭建档的实践研究较少、如何建立相关法律研究不够等问题提出建议。

（8）档案规范。档案规范不仅指对档案内容、格式、管理手续规范化的讨，还包括对规章标准的解读、对比、建议等。如韩雪松[30]从内容确定性和文体风格方面反对用公告进行公示，并就其违规使用问题予以说明，同时对公示性公告与公示性通知之间的互补关系进行叙述；胡明波[31]对《党政机关公文处理工作条例》新旧版本进行比较，新版本体现出鲜明的首创性、简约性、规范性、先进性等特点。

（9）企业档案。企业档案主要是对企业集团从事生产、建设、工程、科研时遇到的问题进行讨论。如徐敏[32]针对工程档案竣工验收过程遇到的各类问题，提出了加大宣传力度、健全管理制度、采取有效措施、加强监督检查的方式来改进；朱艳杰[33]为项目档案的验收，利用问卷调查法详细分析集团基层单位人员配置、项目档案管理方式与信息化建设、保管条件等情况；岳振廷[34]发现加强企业档案文化建设、档案工作主动融入企业文化建设、在档案管理中贯彻以人为本的理念等措施可以加强企业档案管理。

（10）人才培养。人才培养是指高校档案专业学生能力和素质的培育以及档案工作者的继续教育培训。如王广宇[35]基于档案学专业职业核心能力与专业核心能力研判依据、研判维度，提出从人才培养方案的制定、人才培养方式的选择、人才培养资源的构建及人才培养质量的评价角度来培养“双核”人才；王建春[36]根据我国档案专业人员继续教育存在的问题，提出了培育市场力量，扩大继续教育规模；加强教材体系建设，完善继续教育内容；增加实践教学内容，满足一线工作者实际需要；提升师资队伍能力，建设专兼职教师队伍；搭建全国档案教育网络平台，开展远程继续教育；开展公益性培训，控制继续教育收费额度的对策。

（11）电子档案。电子档案以探索保障电子文件真实性、可用性、安全性、完整性的技术、方式等为主。如顾伟[37]针对照片类电子档案部分元数据易篡改的问题，采用电子照片来源检测技术对 M44（设备制造商）和 M45（设备型号）的

chinaXiv:202302.00055v1

真实性进行检测，利用电子照片自身信息的关联性 M57（图像高度）和 M58（图像宽度）的真实性进行检测；王燃[38]通过对电子文件和电子证据的概念、属性进行比较与对接，建议将电子证据的真实性、关联性、合法性及证明力适当吸收至电子文件管理制度中；许晓彤[39]对电子文件“四性”与电子证据“三性”进行系统分析与映射，构建电子文件证据性概念模型。

（12）档案学理论。档案学理论主要以我国档案学理论为研究对象。如李佳男[40]从五个方面对社会记忆可以作为档案学的逻辑起点进行论证；闫静[41]等研究后现代档案学理论的思想实质体现在其理论批判性、思维更新性以及多元主张性；袁也[42]分析文件连续体理论的诞生、发展以及理论成熟三个阶段，认为结构化理论是文件连续体理论诞生的启发者,发展中的参照物以及成果推广的助力者。

（13）档案开发。档案开发是指对档案资源的挖掘以及编研成果的推广，主要包括档案文化产品开发、口述历史编纂、以及多媒体宣传等。任越[43]等分析了档案文化产品“微博”共享模式的运行机理和深广度；张雪[44]在研究故宫文化创意产品的特性、灵魂、定位、宣传与推广方面特点的基础上，建议开发文化创意产品应吸纳民间创作灵感并组建设计开发团队、拓展文创产品的娱乐功能和教育功能、全面改善外部开发环境等；谢兰玉[45]对五大媒体传播口述历史档案信息的优劣势进行总结，发现多方位传播对口述历史档案价值具有重要意义。

（14）档案管理。档案管理以档案收集、整理、移交、库房保管、流转等各业务环节为研究对象，对其出现的问题、原因及对策进行探讨。如姚志刚[46]分析档案库房害虫产生机理以及档案库房霉变产生原理，提出了档案库房虫害与霉变防治的三项原则与三种方法；谢尊贤[47]等为了预防和控制 EMS 寄递高校毕业生档案的流转安全风险，构建了 EMS 寄递高校毕业生档案流转安全风险评价指标体系和多级可拓评价理论模型；苏雅澄[48]以单位名人档案收集工作为例，就其缺少制度、缺乏资金、接收方式单一、收集不完整等问题提出相应的解决对策。

3.3 热点主题

文档-主题概率分布文件是由 D*K 的矩阵组成。如果一篇文档在某个主题中概率越大，则表示文档内容与主题越接近；很多篇文档与某个主题越接近，则表示这个主题是热点主题。根据上述原则，计算热点主题方式如下：

（1）找出每篇文档 d 的主题概率最大值 $\theta_d^{max}$ 并予以标记；

（2）统计每个主题 k 下主题概率最大值 $\theta_d^{max}$ 的个数 $N_k$；

（3）计算主题强度阈值 T，$T = \frac{D}{K}$；

（4）若 $N_k$ >T,则该主题为热点主题。

经计算，热点主题有五个：民国档案、历史档案、档案服务、档案信息化和电子档案。从实际情况来看，档案工作中最重要的是档案服务与档案信息化，档案信息化目前最关注的问题是“增量电子化”，即电子档案的收集与长久保存；从档案作用来看，主要以研究历史为主，即以民国档案和历史档案为主。热点主题计算结果与大众实际关注情况基本相符。

3.4 主题演化分析

chinaXiv:202302.00055v1

将文档-主题概率分布按年度计算，获得各主题在时间窗口上的主题强度演化[49]情况。结果显示 14 个主题中有 3 个主题呈上升趋势，1 个主题呈下降趋势，10 个主题呈不同程度的波段趋势。上升和下降趋势如图 3 所示。

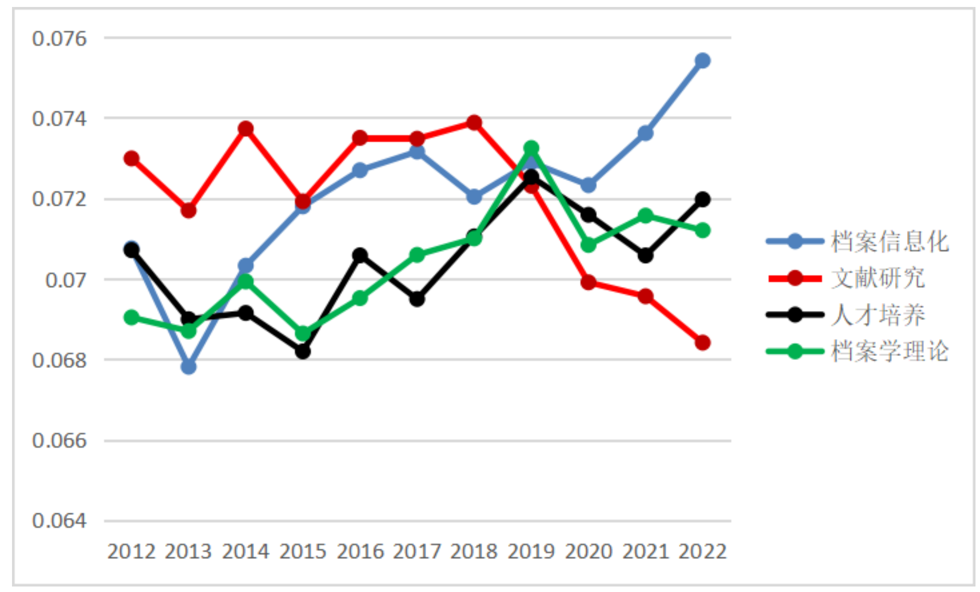



图 3　上升和下降趋势图

文献研究主题为下降趋势，档案信息化、人才培养、档案学理论主题为上升趋势。文献研究主题于 2012-2018 年平稳发展并在 2018 年达到顶峰后开始呈下降趋势，表明其研究趋于成熟。档案信息化主题在 2013 年，2018 年，2020 年有所下降，但很快回升，并逐年增长。档案信息化一直受到学术界重点关注，随着 2020 年国家档案局实施的一系列行业标准后成为最热门的主题之一。人才培养和档案学理论主题虽稍有下降，但总体呈现上升趋势，是未来值得深入研究的对象。对档案学理论研究势必涉及到对档案学理论的宣传教育——人才培养，两者的变化趋势呈现出高度相似性。

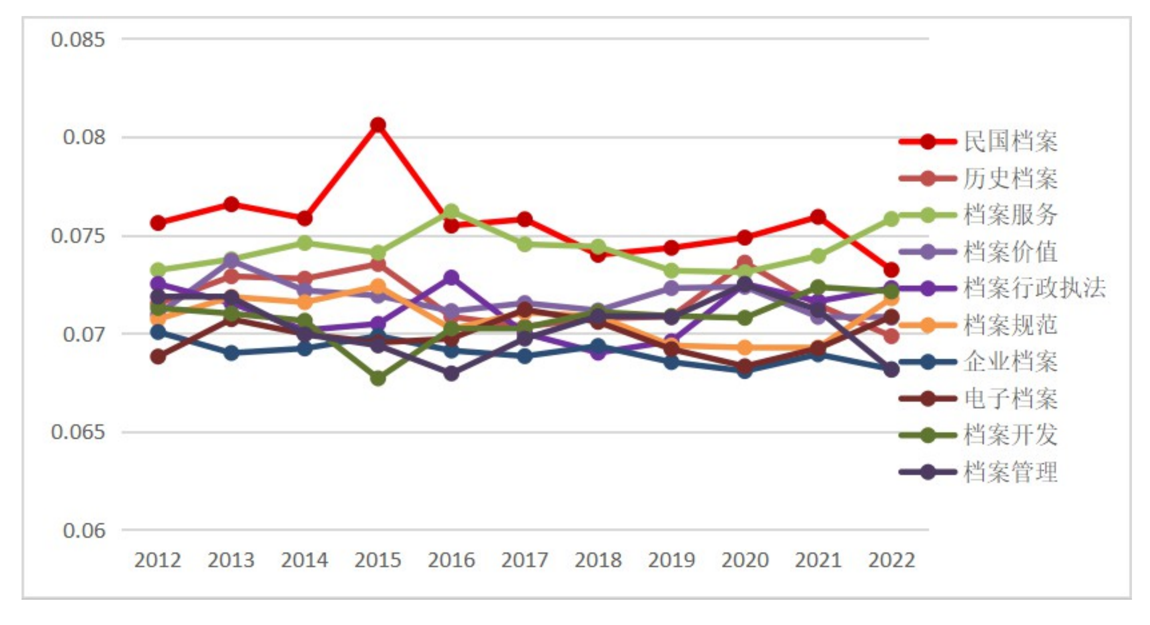



图 4　波段趋势图

波段趋势如图 4 所示。民国档案、历史档案、档案服务、档案价值、档案法律、档案规范、企业档案、电子档案、档案开发、档案管理 10 个主题呈现波段趋势。民

chinaXiv:202302.00055v1

国档案和档案服务主题一直保持较高的关注度。2015年适逢抗战胜利70周年和《新青年》创刊100周年两大历史节点，涌现出一大批学者们从新视角挖掘民国档案主题。历史档案与民国档案主题受关注趋势保持高度一致，但趋势的变化幅度较小，说明学者们对其研究保持一定的持续性。2016年，浙江提出“最多跑一次”改革，档案服务主题达到短暂顶峰后幅度有所下降，于2022年习近平总书记对新时代档案工作重要指示后达到第二个顶峰。档案服务与档案开发主题表现出较为相反的幅度变化。档案服务是及时为用户提供所需的档案信息，是被动行为；档案开发是档案专业人士通过筛选、编纂等方式加工档案信息，是主动行为。两者相互衔接、互相补充，为用户更好地利用档案信息提供便利。档案价值主题于2012-2013年稍有上升，随后开始下降趋于平稳。说明该主题经过前期发展现处于平稳发展阶段。档案法于2016年、2020年进行修订，档案法律主题强度随之达到新高，受到学者们高度重视。档案法律和档案规范主题呈现出相反的变化趋势。两者相辅相成，共同维护档案安全，为档案事业发展提供指引、教育、强制作用。企业档案主题在2012-2022年间虽有所变化，但变化很小，趋于平稳，受到各企业研究者的连续关注。电子档案主题一直是学者们重点关注对象，在2017年国家档案局实施的《电子文件归档与电子档案管理规范》文件时关注度达到极限后出现发展瓶颈，研究成果大幅减少直至2020年档案法中提出“电子档案与传统载体档案具有同等效力”等一系列支持电子档案单套制管理的原则后，其发展开始加速回升。档案管理主题于2012-2016年、2020-2022年呈下降趋势，2016-2020年呈上升趋势，总体呈现波段变化。档案管理随着新法规、新政策、新技术的提出而不断发生变革，主题强度变化幅度较大。

4 结论

本文通过gensim工具包搭建LDA模型，对2012年-2022年间9本档案学核心期刊进行数据挖掘。根据pyLDAvis工具包的可视化结果为主题命名，根据文档一主题概率分布情况并结合时间项分析热点主题和主题演化过程。分析结果与实际研究情况高度匹配，验证了LDA模型在档案学文献研究的有效性。从不同维度展示了近十一年档案学领域的研究成果，不同主题下的概率较高的文档有助于对某些特定内容进行精细化分析。希望为档案学者们的后续研究提供参考价值

**参考文献:**


[1] 宋进.河南省公共档案馆学术创新能力评价研究[J].档案管理,2021,(04):91+93.

[2] 马双双,齐俊景,韩彤彤.我国档案学领域研究机构合作网络分析——以档案学7种核心期刊为例[J].档案与建设,2021,(01):37-43+11.

[3] 李英.我国图书情报与档案管理学科研究现状剖析——基于2009-2013年国家自然科学基金和国家社会科学基金立项的分析[J].图书情报工作,2014,58(09):31-36.

[4] 陶俊,黄新荣.高被引主题结构与档案学科竞争力[J].图书情报工作,2019,63(16):14-21.

[5] 杨万欢,向禹.国内智慧档案馆研究动态分析——基于核心期刊和CSSCI(2013—2019年)的文献计量研究[J].档案管理,2020,(04):98-99.

[6] 于欢欢,程慧平.区块链技术在国内电子档案管理中的应用研究述评[J].档案与建设,2021(05):27-33.


chinaXiv:202302.00055v1

chinaXiv:202302.00055v1

[7] 张晓培.基于网络的政府信息公开与档案相关研究文献关键词词频分析[J].档案管理,2017,(05):63-64.

[8] 逯万辉,荆林波.基于作者主题模型的学者聚类与学术影响力评价方法研究[J].情报资料工作,2020,41(04):60-66.

[9] 宋雪雁,张伟民,张祥青.基于档案文献的清代祭祀礼器知识图谱构建研究[J].图书情报工作,2022,66(03):140-151.

[10] 马海群,张涛.基于文本计算的我国档案政策法律协同性研究——以《中华人民共和国档案法》（2020 修订版）和《“十四五”全国档案事业发展规划》为蓝本[J].档案学研究,2022,(02):26-32.

[11] Griffiths T L, Steyvers M. Finding scientific topics[J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101 Suppl 1(1): 5228-5235.

[12] Luo R , Xu J , Zhang Y , et al. PKUSEG: A Toolkit for Multi-Domain Chinese Word Segmentation:, 10.48550/arXiv.1906.11455[P]. 2019.

[13] 张东鑫,张敏.图情领域 LDA 主题模型应用研究进展述评[J].图书情报知识:1-14.

[14] Sievert C , Shirley K E . LDAvis: A method for visualizing and interpreting topics[C]// Workshop on Interactive Language Learning, Visualization, and Interfaces at the Association for Computational Linguistics. 2014.

[15] 张展.“元首”之争——华北、华中日军围绕“中央政权”的博弈[J].民国档案,2020,No.140(02):93-105.

[16] 姚江鸿.军事冲击下的政治改革——1944 年国民党对政治改革的考量和内部决策[J].民国档案,2021,No.146(04):131-142

[17] 李兵.清代两湖南北分闱再探[J].历史档案,2013(01):78-84.

[18] 陈晨.清代盛京巡察官考述[J].历史档案,2018,No.151(03):85-91.

[19] 赵彦昌,王睿嘉.《黑图档》中的盛京皇宫凤凰楼[J].北京档案,2021,No.369(09):46-48.

[20] 王成兴,许炎.档案信息服务交互平台建设初探[J].北京档案,2019,No.338(02):7-10.

[21] 傅永珍.面向用户需求的个性化档案信息服务探讨[J].北京档案,2013,No.267(03):39-41.

[22] 连志英,朱宏涛.参与式档案信息服务模式:社会化媒体环境下档案信息服务新模式[J].档案学通讯,2018,(04):59-64.

[23] 赵爱学.档案文化溯源的一种考究[J].档案学通讯,2013,(04):17-20.

[24] 杨光,奕窕.记录媒介演进与档案历史叙事的变迁[J].档案学通讯,2019,No.248(04):19-27.

[25] 管先海.对档案价值、档案主体价值与档案客体价值的认识——兼与归吉官先生商榷[J].档案管理,2016,No.220(03):11-13.

[26] 梅帅.论新《行政处罚法》视域下档案行政处罚程序的完善[J].档案学通讯,2022,(04):67-75.

[27] 金波,杨鹏.大数据时代档案数据安全保障探究[J].档案学通讯,2022,(03):30-38.

[28] 赵湘渝.智能化技术在档案信息化建设中的应用现状与对策分析[J].档案管理,2022,No.256(03):73-74.

[29] 王国强.1985 年-2018 年我国家庭建档研究文献统计分析[J].档案管理,2018,(04):71-72.

[30] 韩雪松.试析公示性公告的文体理据及行文问题[J].档案管理,2017,(03):76-78.

[31] 胡明波.《党政机关公文处理工作条例》特点研究[J].档案学通讯,2012,No.209(05):11-14.

[32] 徐敏.中小型水利工程竣工验收中档案管理存在的问题及对策[J].档案管理,2022,(04):121-122.

[33] 朱艳杰.中国平煤神马集团项目档案基础管理及专项验收情况调查分析[J].档案管理,2016,No.223(06):65-66.

[34] 岳振廷.企业档案管理推动企业文化建设路径探析[J].浙江档案,2018,No.450(10):62-63.

[35] 王广宇.论档案学专业的“双核”能力与培养[J].档案管理,2022,(04):96-98.

[36] 王建春,张凤霞.我国档案专业人员继续教育存在的问题与解决对策[J].浙江档案,2019,No.462(10):57-59.

[37] 顾伟.照片类电子档案元数据真实性研究[J].档案学研究,2022,(01):92-96.

[38] 王燃.电子文件管理与证据法规则的契合研究[J].档案学通讯,2018,No.243(05):51-56.

[39] 许晓彤.电子文件证据性概念模型研究[J].档案管理,2021,No.249(02):27-31.

[40] 李佳男.社会记忆——档案学逻辑起点探究[J].档案管理,2022,(05):18-22+26.

[41] 闫静,徐拥军.后现代档案学理论的思想实质研究[J].档案学研究,2019,No.169(04):4-12.

[42] 袁也.结构化理论对文件连续体理论影响探析[J].档案学通讯,2016,No.230(04):35-39.

[43] 任越,曹玉.以“微博”为平台的档案文化产品共享现状分析与发展策略研究[J].档案管理,2014,No.208(03):27-29.

chinaXiv:202302.00055v1

[44] 张雪.故宫博物院文化创意产品开发及启示[J].浙江档案,2019,No.454(02):20-21.

[45] 谢兰玉.口述历史档案价值实现的媒体路径研究[J].浙江档案,2014,No.403(11):18-19.

[46] 姚志刚.档案库房虫害与霉变防治研究[J].北京档案,2020(09):25-28.

[47] 谢尊贤,李艳艳,文小琼,杨彬,米顺.基于物元多级可拓模型的档案流转安全风险评价研究[J].档案学研究,2018,No.164(05):119-124.

[48] 苏雅澄.高校名人档案收集工作的策略研究[J].北京档案,2016,No.312(12):26-27.

[49] 林丽丽,马秀峰.基于LDA模型的国内图书情报学研究主题发现及演化分析[J].情报科学,2019,37(12):87-92.

**作者贡献说明:**周洁:文献搜集，算法实现与论文撰写等；盛梅:提供思路，论文修改等。

**Hot Topics and Evolution Research of Archives in China Based on LDA Model***

Zhou jie[1] , Sheng mei[2]

[1]Ningbo University Archives  Ningbo  315211 [2]Tongde Hospital of Zhejiang Province  Hangzhou  310000

**Abstract:** **[Purpose/significance]** The LDA model is used to discover the research hotspots and development trends of Chinese archival science in recent years, which provides data support and reference value for the research of Chinese archival science.
**[Method/process]** Select the Chinese abstracts of 9 core journals of archival science from 2012 to 2022 as the analysis sample, take the CNKI database as the source database, use the Python open source toolkit pkuseg to carry out Chinese word segmentation, build the LDA model with  gensim, and use pyLDAvis presents interactive visualization of various topics based on web. According to the visualization results of pyLDAvis, the topic is named. According to the probability distribution of document-topic and the time item, the hot topics and the topic evolution process are analyzed.
**[Result/conclusion]** According to the LDA model, it is possible to effectively distinguish the topics of research in the field of archival science in China. From 2012 to 2022, there are 14 topics in the field of archival science in China, including 5 hot topics; Three topics showed an upward trend, one topic showed a downward trend, and 10 topics showed a band trend of varying degrees.

**Keywords:**Archives   LDA model   Hot Topics   Topic Evolution

chinaXiv:202302.00055v1